HHS Public Access

Author manuscript

Biometrics. Author manuscript; available in PMC 2015 July 31.

Published in final edited form as:

Biometrics. 2014 September ; 70(3): 762–770. doi:10.1111/biom.12186.

Partial Correlation Matrix Estimation using Ridge Penalty Followed by Thresholding and Reestimation

Min Jin Ha and

Department of Biostatistics, UNC Chapel Hill, North Carolina, U.S.A.

Wei Sun

Department of Biostatistics, Department of Genetics, UNC Chapel Hill, North Carolina, U.S.A.

Min Jin Ha: mjha@live.unc.edu; Wei Sun: weisun@email.unc.edu

Summary

Motivated by the problem of construction gene co-expression network, we propose a statistical framework for estimating high-dimensional partial correlation matrix by a three-step approach. We first obtain a penalized estimate of a partial correlation matrix using ridge penalty. Next we select the non-zero entries of the partial correlation matrix by hypothesis testing. Finally we reestimate the partial correlation coefficients at these non-zero entries. In the second step, the null distribution of the test statistics derived from penalized partial correlation estimates has not been established. We address this challenge by estimating the null distribution from the empirical distribution of the test statistics of all the penalized partial correlation estimates. Extensive simulation studies demonstrate the good performance of our method. Application on a yeast cell cycle gene expression data shows that our method delivers better predictions of the protein-protein interactions than the Graphic Lasso.

Keywords

Co-expression network; Empirical null distribution; Graphical model; Partial correlation matrix; Ridge regression

1. Introduction

The expression of multiple genes can be studied through a network perspective, where the set of genes of interest are vertices and the relations among the genes are undirected/directed edges. The gene co-expression network analysis is a popular approach to dissect gene expression regulation patterns and to detect functionally related genes (Stuart et al., 2003; de Jong et al., 2012). In this paper we study the (undirected) co-expression network of a group of genes constructed through their partial correlation matrix.

Correspondence to: Min Jin Ha, `mjha@live.unc.edu`; Wei Sun, `weisun@email.unc.edu`.

Supplementary Materials

Web Appendices, Tables, Figures referenced in Sections 1, 2.2., 2.4, 3.3, and 3.4 and the `GGMridge` R package are available with this paper at the *Biometrics* website on Wiley Online Library.

We denote the expression of p genes by a p-dimensional random vector: $\boldsymbol{X} = (X_1, \ldots, X_p)^{\mathrm{T}} \in \mathbb{R}^p$ with an unknown, but positive definite covariance matrix $\boldsymbol{\Sigma}$. Let $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ be the inverse of the covariance matrix $\boldsymbol{\Sigma}$, with its element at a-th row and b-th column denoted by $\boldsymbol{\Omega}_{ab}$. $\boldsymbol{\Omega}$ is also called concentration matrix or precision matrix. The partial correlation between X_a and X_b is a measure of the linear relationship between X_a and X_b after accounting for the linear effects of all the remaining variables (Christensen, 2002). The partial correlations can be obtained by the off diagonal elements of the negative definite matrix $-\mathtt{scale}(\boldsymbol{\Omega})$:

$$\mathbf{R} = [\rho_{ab}]_{p\times p} = -\,\mathtt{scale}(\boldsymbol{\Omega}), \quad (1)$$

where the $\mathtt{scale}$ is an operator defined for a square matrix. Let $\mathtt{diag}(\mathbf{A})$ be a diagonal matrix constructed by the diagonal elements of $\mathbf{A}$, then $\mathtt{scale}(\mathbf{A}) = \mathtt{diag}(\mathbf{A})^{-1/2}\mathbf{A}\mathtt{diag}(\mathbf{A})^{-1/2}$. The derivation of equation (1) is presented in the Section A of the Supplementary Materials. The zero structure of the partial correlation matrix of p random variables can be represented by an undirected graph

$$\mathbf{G} = (\boldsymbol{\Gamma}, \mathbf{E}),$$

where $\boldsymbol{\Gamma} = \{1, \cdots, p\}$ is the set of vertices and $\mathbf{E}$ is a set of edges in $\boldsymbol{\Gamma} \times \boldsymbol{\Gamma}$ such that any edge between vertices a and b belongs to $\mathbf{E}$ if and only if $\rho_{ab} \neq 0$. We refer to such an undirected graph $\mathbf{G}$ as a *partial correlation graph*. Under multivariate Gaussian distribution assumption for $\boldsymbol{X}$, zero partial correlation between two variables is equivalent to the conditional independence of these two variables given the remaining variables.

Although many methods have been developed for partial correlation matrix estimation in high dimensional problems where $p > n$, we find that a simple penalized estimation using ridge penalty has favorable properties. The advantage of this ridge penalization approach has not been appreciated in the existing literature, partly because it does not provide sparse estimates, i.e., none of the partial correlation is estimated exactly as 0. We propose a novel approach to threshold the ridge estimates to select non-zero entries of the partial correlation matrix. Finally we reestimate the partial correlation coefficients on the none-zero entries of the partial correlation matrix. Thresholding ridge estimates is desirable because it leads to parsimonious and more interpretable partial correlation matrix estimate, and further reduces estimation error as well.

Next we briefly review the existing works for estimating concentration matrix or partial correlation matrix and related statistical inference. Suppose there are n independent samples of p random variables $\boldsymbol{X} = (X_1, \ldots, X_p)^{\mathrm{T}}$, and let $\mathbf{X}$ be the $p \times n$ data matrix. Schäfer et al. (2005) proposed to estimate covariance matrix by

$$\hat{\mathbf{S}} = (1-\lambda)\mathbf{S} + \lambda\mathbf{T}, \quad (2)$$

where $\mathbf{S}$ is sample covariance matrix, and $\mathbf{T}$ is a target matrix. Schäfer et al. (2005) derived optimal choice of λ to minimize squared loss of covariance matrix estimate for six commonly used target matrices $\mathbf{T}$'s. Then they propose to estimate partial correlation matrix by the inverse of their correlation matrix estimate. Assume $\boldsymbol{X} = (X_1, \ldots, X_p)^{\mathrm{T}}$ follows

multivariate Gaussian distribution with zero mean, denoted by $N(\mathbf{0}, \mathbf{\Sigma})$. The log-likelihood of concentration matrix $\mathbf{\Omega}$ is proportional to

$$l(\mathbf{\Omega}) = \log|\mathbf{\Omega}| - \mathtt{tr}(\mathbf{S\Omega}), \quad (3)$$

where $\mathtt{tr}(\cdot)$ is trace of a square matrix and $\mathbf{S}$ is the sample covariance matrix. When $n \geq p$, $\mathbf{S}$ is positive definite with probability 1 and $\mathbf{S}^{-1}$ is the maximum likelihood estimate (MLE) of $\mathbf{\Omega}$ (Lauritzen, 1996). However, this approach fails when $p > n$, and may perform poorly unless n is much larger than p. Therefore MLE with certain constraints or penalized MLE are often used for high dimensional problems when p is lager or much larger than n. Examples include covariance selection from positive definite matrices (Dempster, 1972) or iterative partial maximization based on deviance tests (Speed and Kiiveri, 1986). More general linear restrictions on edges are enabled by colored graph models (Højsgaard and Lauritzen, 2008). Recently, many penalized MLE of $\mathbf{\Omega}$ have been proposed for high dimensional problems (Yuan and Lin, 2007; Rothman et al., 2008; Banerjee et al., 2008; Friedman et al., 2008; Fan et al., 2009). One of the most widely used methods is the graphic Lasso (Friedman et al., 2008), which maximizes the following penalized log likelihood:

$$l(\mathbf{\Omega}) = \log|\mathbf{\Omega}| - \mathtt{tr}(\mathbf{S\Omega}) - \kappa \sum_{\mathtt{a},\mathtt{b}} |\mathbf{\Omega}_{\mathtt{ab}}|, \quad (4)$$

where κ is a tuning parameter.

With a focus on determining the partial correlation graph, rather than precise estimation of the partial correlation coefficients, Meinshausen and Bühlmann (2006) proposed a regression-based approach called *neighborhood selection*. The neighborhood for each vertex was estimated by penalized regression of the corresponding variable versus the remaining variables. Banerjee et al. (2008); Friedman et al. (2008) showed that estimating the penalized MLE (with L_1 penalty) of $\mathbf{\Omega}$ could be viewed as p-coupled iterative versions of the p separate neighborhood selections. More recent methodology developments related with neighborhood selection include Yuan (2010) and Zhou et al. (2011).

Statistical inference of partial correlation estimates is another topic related with our method development, particularly the second step of our method for thresholding partial correlations. Given a partial correlation estimate, denoted by $\hat{\rho}$, one may test $H_0 : \rho = 0$ against $H_A : \rho \neq 0$ using a test statistic constructed by Fisher's Z-transformation: $\psi(\hat{\rho}) = 0.5 \log\{(1 + \hat{\rho}) / (1 - \hat{\rho})\}$. Specifically, one may reject the null hypothesis at level α if $(n - p - 1)^{1/2}|\psi(\hat{\rho})| > \Phi^{-1}(1 - \alpha/2)$ for standard normal c.d.f. Φ (Anderson, 2003). However, this testing procedure assumes the sample size n is substantially greater than p. For high dimensional problems with $p > n$, Schäfer and Strimmer (2005) proposed an inference approach by assuming partial correlation estimates across all variables followed a mixture of null and alternative distributions where the null was Hotelling distribution with unknown degree of freedom and the alternative was uniform (−1,1). Magwene et al. (2004) and Wille et al. (2004) proposed to use low-order partial correlations to avoid singularity problem when $p > n$. Subsequently, Wille and Bühlmann (2006) discussed more formal statements on Gaussian graphical model inference using low-order partial correlations. Castelo and

Roverato (2006) generalized the 0–1 partial correlation graph to an arbitrary $q \leq p - 2$ order partial correlation graph.

The remaining parts of the paper are organized as follows. We present our method in Section 2, demonstrate the effectiveness of our method by simulations and real data analysis in Section 3, and conclude this paper by some discussions in Section 4.

2. Method

2.1 Estimation of partial correlation matrix using ridge penalty

Without loss of generality, we assume each row of the $p \times n$ data matrix $\mathbf{X}$ has been standardized to have mean 0 and standard deviation 1 so that $\mathbf{S} = \mathbf{X}\mathbf{X}^{\mathrm{T}}/n$ is the sample correlation matrix. Then a straightforward estimate of the off-diagonal elements of a partial correlation matrix can be obtained from

$$\hat{\mathbf{R}} = -\,\mathtt{scale}(\mathbf{S}^{-1}).$$

However, when $n < p$, $\mathbf{S}$ is not invertible. To solve the singularity problem of inverting a sample correlation matrix, we add a positive constant to the diagonal elements of the sample correlation matrix:

$$\hat{\mathbf{R}}(\lambda) = -\,\mathtt{scale}((\mathbf{S} + \lambda \mathbf{I}_{\mathrm{p}})^{-1}), \quad (5)$$

where $\lambda \geq 0$ and $\mathbf{I}_{\mathrm{p}}$ is a $p \times p$ identity matrix. We call $\mathbf{S}^{+}(\lambda) = (\mathbf{S} + \lambda \mathbf{I}_{p})^{-1}$ as the *ridge* inverse in the analogy to ridge regression (Hoerl and Kennard, 1970). The modified sample covariance matrix $\mathbf{S} + \lambda \mathbf{I}_{p}$ guarantees full rank for any $\lambda > 0$, and has been used as an initial covariance matrix estimate in the coordinate descent algorithms in Banerjee et al. (2008) and Friedman et al. (2008).

Next we show that as λ varies from 0 to ∞, $\hat{\mathbf{R}}(\lambda)$ varies from a scaled generalized inverse to an identity matrix. Let $\mathbf{X}/\sqrt{n} = \mathbf{U}\mathbf{D}\mathbf{V}^{\mathrm{T}}$ be a singular value decomposition with $\mathtt{rank}(\mathbf{X}) = k \leq \mathtt{min}(n, p)$, where $\mathbf{U}$ and $\mathbf{V}$ are, respectively $p \times p$ and $n \times n$ orthogonal matrices, $\mathbf{D}$ is $p \times n$ diagonal matrix with its first k nonzero diagonal elements $d_1, \ldots, d_k$ and all other elements being zero. Since $\mathbf{S}^{+}(\lambda) = \mathbf{U}(\mathbf{D} + \lambda \mathbf{I}_{p})^{-1}\mathbf{U}^{\mathrm{T}}$, it is obvious that

$$\mathtt{scale}(\mathbf{S}^{+}(\lambda)) \to \mathtt{scale}(\mathbf{S}^{-}) \mathrm{as} \lambda \to 0, \quad (6)$$

where $\mathbf{S}^{-}$ is Moore-Penrose generalized (MPG) inverse of $\mathbf{S}$ if $k < p$ (Schott, 2005). By the invariance of the `scale` operator under scalar product,

$$\mathtt{scale}(\mathbf{S}^{+}(\lambda)) = \mathtt{scale}(\lambda \mathbf{S}^{+}(\lambda)) \to \mathbf{I}_{\mathrm{p}} \mathrm{as} \lambda \to \infty. \quad (7)$$

Since the estimates of regression coefficients using MPG inverse is minimum L_2 solution (proposition 1 of Lv and Fan (2009)), $\mathbf{S}^{+}(\lambda)$ goes to k rank ridge inverse when λ goes to 0 by (6). From (7) the partial correlation matrix shrinks toward the identity matrix as λ goes to

infinity. In practice, the optimal performance of this ridge estimate relies on an appropriate choice of λ, which will be addressed after we introducing the other two steps of our method.

2.2 Thesholding

We propose a hypothesis testing approach to threshold the ridge estimate of partial correlations $\hat{\mathbf{R}}(\lambda)=[\hat{\rho}_{ab}^{\lambda}]_{p\times p}$, where λ is the tuning parameter for ridge estimate. We first apply Fisher's Z-transformation on partial correlations estimate, denoted by $\{\psi(\hat{\rho}_{ab}^{\lambda}): a \in \mathbf{\Gamma}, b \in \mathbf{\Gamma}, \text{ and } a \neq b\}$. We assume these z-statistics follow a mixture of null and alternative distributions, corresponding to the cases where the true partial correlations are zero or non-zero, respectively. By assuming the vast majority of the z-statistics in the central part of this mixture distribution (i.e., a region around 0) arise from the null distribution, we estimate the null distribution using Efron's central matching method (Efron, 2004). Specifically, we assume the observed z-statistics follow a *mixture* distribution

$$f(\psi)=\eta f_0(\psi)+(1-\eta)f_a(\psi), \quad (8)$$

where the null distribution $f_0(\psi)$ is a normal distribution $N(\mu_0,\sigma_0^2)$, the alternative distribution $f_a(\psi)$ is left un-specified, and η is the proportion of the z-statistics arising from the null distribution. Using Efron's central matching method (Efron, 2004), we estimate the null distribution $N(\mu_0,\sigma_0^2)$ by matching the mixture distribution and the null distribution at the central part of the distributions. Specifically, assuming $f(\psi) = f_0(\psi)$ around $\psi = 0$ gives

$$\log f(\psi)=-\frac{1}{2}\left(\frac{\psi-\mu_0}{\sigma_0}\right)^2+C \quad (9)$$

for a constant *C*.

We estimate $f(\psi)$ (the density of the mixture distribution) using polynomial Poisson regression. The range of the $p(p-1)/2$ observed ψ values is partitioned into *K* equal intervals with interval *k* having mid point x_k and s_k observed ψ values. s_k's (k=1,…,K) are assumed to be independently distributed following Poisson distributions with mean ν_k's. We fit a *q* degree polynomial Poisson regression on ν_k,

$$\log(\nu_k)=\log(f(x_k)/c)=\sum_{j=1}^{q}\theta_j(x_k)^j, \quad (10)$$

for $k = 1, \ldots, K$ and a normalizing constant *c* making the marginal density $f(\psi)$ integrated to 1. The estimates of $\{\theta_j : j = 1, \ldots, q\}$ are used to estimate $\log(\hat{f}(\psi)/c)=\sum_{j=1}^{q}\hat{\theta}_j\psi^j$. Then using equation (9), we can obtain the estimates of μ_0 and σ_0:

$$\hat{\mu}_0=\arg\max_{\psi}\left\{\hat{f}(\psi)\right\}, \text{and}\hat{\sigma}_0=\left[-\frac{d^2}{d\psi^2}\log\hat{f}(\psi)\right]_{\psi=\hat{\mu}_0}^{-\frac{1}{2}}. \quad (11)$$

With the estimate of null distribution $N(\hat{\mu_0}, \hat{\sigma_0})$, we can calculate p-values for each partial correlation estimate. The degree of polynomial regression, q, is a nuisance parameter. Based on the sparsity assumption that most p-values arise from the null, we choose the q so that the p-values are most uniformly distributed. The empirical distribution function of the p-values, $\{\pi_{ab}^{(q)} | a \neq b \in \mathbf{\Gamma}\}$ given q, is

$$F_q(\pi) = \frac{2}{p(p-1)} \sum_{a,b \in \mathbf{\Gamma}, a \neq b} I(\pi_{ab}^{(q)} \leq \pi). \quad (12)$$

We suggest to estimate q by

$$\hat{q} = \arg \min_q \left[\sup_{0<\pi<1} |F_q(\pi) - F_0(\pi)| \right], \quad (13)$$

where $F_0(\pi)$ is uniform distribution between 0 and 1, and $D_q = \sup_{0<\pi<1} |F_q(\pi) - F_0(\pi)|$ is a distance measure used in Kolmogorov-Smirnov statistic. Figure S6 (Supplementary Materials Section F) displays the average and one standard deviation of D_q values over 100 simulation data sets for $p = 500$, $n = 30$ and $\eta = 1$ or $\eta = 0.9997$, which corresponds to 38 non-zero partial correlations. Adding 38 nonzero partial correlations to the null needs 3 or 4 higher polynomial order on average to estimate the null distribution. Finally, a threshold α is needed to select non-zero entries of the partial correlation matrix. We select α by cross-validation, and we defer the discussion of details to section 2.4. Given this threshold, we can estimate the sparsity η. An upper-bound of η can also be estimated following (Efron, 2004) (Supplementary Materials Section C). From our simulations, the estimate of η based on our cross-validation selected threshold is more accurate.

2.3 Re-estimation of partial correlation coefficients

Given the partial correlation graph structure estimated in the previous step, we re-estimate the partial correlation coefficients at the non-zero entries of the partial correlation matrix. Suppose that the covariance matrix $\mathbf{\Sigma}$ and the concentration matrix $\mathbf{\Omega}$ are partitioned according to random variables X_a and $\boldsymbol{X}_{-a}$ where $\boldsymbol{X}_{-a}$ is a $(p-1) \times 1$ random vector except for a random variable X_a. The blocks are denoted by $\mathbf{\Sigma}_{a,a}$, $\mathbf{\Sigma}_{-a,a}$, $\mathbf{\Sigma}_{a,-a}$, $\mathbf{\Sigma}_{-a,-a}$ and $\mathbf{\Omega}_{a,a}$, $\mathbf{\Omega}_{-a,a}$, $\mathbf{\Omega}_{a,-a}$, $\mathbf{\Omega}_{-a,-a}$. Consider the best linear predictor of X_a by $\boldsymbol{X}_{-a}^{\mathrm{T}} \boldsymbol{\beta}^a$ for any $a \in \mathbf{\Gamma}$. Let $\varepsilon_a = X_a - \boldsymbol{X}_{-a}^{\mathrm{T}} \boldsymbol{\beta}^a$. It is easy to show that $\boldsymbol{\beta}^a = \mathbf{\Sigma}_{-a,-a}^{-1} \mathbf{\Sigma}_{-a,a}$ and $\mathtt{Var}(\mathtt{X_a} - \boldsymbol{X}_{-\mathtt{a}}^{\mathrm{T}} \boldsymbol{\beta}^{\mathtt{a}}) = \mathtt{Var}(\varepsilon_{\mathtt{a}}) = \mathbf{\Sigma}_{\mathtt{a},\mathtt{a}} - \mathbf{\Sigma}_{\mathtt{a},-\mathtt{a}} \mathbf{\Sigma}_{-\mathtt{a},-\mathtt{a}}^{-1} \mathbf{\Sigma}_{-\mathtt{a},\mathtt{a}}$. From inverse formula for block matrix and $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$,

$$\mathbf{\Omega}_{a,a} = (\mathtt{Var}(\varepsilon_{\mathtt{a}}))^{-1}, \mathbf{\Omega}_{-a,a} = -(\mathtt{Var}(\varepsilon_{\mathtt{a}}))^{-1} \boldsymbol{\beta}^a. \quad (14)$$

From $\hat{\mathbf{E}}(\lambda, \alpha)$ estimated in the thresholding step, we know all the variables adjacent to $a \in \mathbf{\Gamma}$, denoted by $\hat{\mathtt{ne}}_a$. Based on the sparsity assumption, we assume $|\hat{\mathtt{ne}}_a| < n$, then we can have the following refined estimates of the concentration matrix:

$$\hat{\boldsymbol{\Omega}}_{a,a}=(n-|\hat{\mathtt{ne}}_a|)/\|\mathbf{X}_a-\mathbf{X}_{\hat{\mathtt{ne}}_a}\hat{\beta}^{a,\hat{\mathtt{ne}}_a}\|_2^2, \tilde{\boldsymbol{\Omega}}_{-a,a}=-\tilde{\boldsymbol{\Omega}}_{a,a}\hat{\beta}^{a,\hat{\mathtt{ne}}_a}, \quad (15)$$

where $\hat{\beta}^{a,\hat{\mathtt{ne}}_a}=(\mathbf{X}_{\hat{\mathtt{ne}}_a}\mathbf{X}_{\hat{\mathtt{ne}}_a}^{\mathrm{T}})^{-1}\mathbf{X}_{\hat{\mathtt{ne}}_a}\mathbf{X}_a$, $\mathbf{X}_{\hat{\mathtt{ne}}_a}$ is $|\hat{\mathtt{ne}}_a|\times n$ submatrix of $\mathbf{X}$ corresponding to $\hat{\mathtt{ne}}_a$. Since this solution is not symmetric in general, we set the final estimates of the off-diagonal elements of $\boldsymbol{\Omega}$ as

$$\hat{\boldsymbol{\Omega}}_{a,b}=\hat{\boldsymbol{\Omega}}_{b,a}=\mathtt{sign}(\hat{\beta}_{\mathtt{b}}^{\mathtt{a},\hat{\mathtt{ne}}_{\mathtt{a}}})\sqrt{|\tilde{\boldsymbol{\Omega}}_{a,b}\tilde{\boldsymbol{\Omega}}_{b,a}|}\,\mathrm{for}\,\mathtt{a}\neq\mathtt{b}, \quad (16)$$

and obtain the estimate the partial correlation coefficients from $-\mathtt{scale}(\hat{\boldsymbol{\Omega}})$.

2.4 Tuning parameter selection

Our method has two tuning parameters. One is λ for ridge estimate of partial correlation matrix and the other is α, the p-value cutoff. We perform a two-grid search of the combination of λ and α by minimizing sum squared prediction errors of *p* separate ridge regressions. Specifically, given λ and α, let $N_a\equiv\hat{\mathtt{ne}}_a$ be the neighborhood of X_a after thresholding. Let $\mathbf{X}_a$ be an $n\times 1$ vector of the measurements of variable X_a, and let $\mathbf{X}_{N_a}$ be an $n\times|\hat{\mathtt{ne}}_a|$ matrix of the measurements of the variables $\mathbf{X}_{N_a}=\{X_k:k\in N_a\}$. Then the sum squared prediction error is

$$\sum_a[\mathbf{X}_a-\mathbf{X}_{N_a}\hat{\beta}^a(\lambda)]^{\mathrm{T}}[\mathbf{X}_a-\mathbf{X}_{N_a}\hat{\beta}^a(\lambda)],$$

where

$$\hat{\beta}^a(\lambda)=\arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{p-1}}\frac{1}{n}(\mathbf{X}_a-\mathbf{X}_{N_a}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{X}_a-\mathbf{X}_{N_a}\boldsymbol{\beta}).$$

We present the details of this tuning parameter selection method and its justifications into Supplementary Materials Section B.

As an alternative, one may conduct *p* separate ridge regressions with different λ's to estimate a partial correlation matrix. This is equivalent to a general form of the ridge estimation of partial correlation matrix:

$$\hat{\mathbf{R}}(\lambda)=-\mathtt{scale}((\mathbf{S}+\mathtt{diag}(\lambda_1,\ldots,\lambda_{\mathtt{p}}))^{-1}),$$

where $\lambda_1>0,\ldots,\lambda_p>0$ and $\mathtt{diag}(\lambda_1,\ldots,\lambda_p)$ is a $p\times p$ diagonal matrix with diagonal elements $\lambda_1,\ldots,\lambda p$. For all $i=1,\ldots,p$, we choose the tuning parameter λ_i by minimizing 10-fold cross validation estimates of the prediction errors of the ridge regression with X_i as response variable and all the other variables as covariates. We refer this method as *separate estimation*, and refer the method with $\lambda=\lambda_1=\ldots=\lambda_p$ as *joint estimation*. We show that the joint estimation, although suffers from larger bias, does provide more stable ordering of the

partial correlation coefficients than the separate estimate, and hence more accurate estimate of partial correlation matrix after thresholding and re-estimation, which eliminate the bias of initial estimates (Supplementary Materials, Section D).

3. Results

3.1 Simulation I

We first use a simple simulation to demonstrate that the p-values calculated using central matching method follow the expected uniform distribution under null, while the p-values calculated using asymptotic distribution can lead to inflated type I error. We simulated data from multivariate Gaussian distribution $N(\mathbf{0}, \mathbf{I}_{p\times p})$ with $p = 100$ and $n = 1000$ or 110. All pairwise partial correlations were calculated by inverting the sample correlation matrix, and then the test statistics were calculated by Fisher's Z transformation of the partial correlations. The p-values of the test statistics were calculated using theoretical null distribution $N(0, 1/n - p - 1)$, and the empirical null distribution estimated by central matching method. As shown in the qq-plots of Figure 1, when $p = 100$ and $n = 1000$, p-values calculated using either the theoretical null distribution or the empirical null distribution followed the expected uniform distribution. However when the sample size was decreased to $n = 110$, the p-values calculated from the empirical null distribution were still uniformly distributed but the p-values calculated from the theoretical null distribution were severely inflated.

3.2 Simulation II

We consider random networks where both the network structure and the partial correlation coefficients are random. The only restriction is that the partial correlation matrix is diagonally dominant, so that **R** is a strictly negative definite matrix. The simulation datasets were generated following similar approach of Schäfer and Strimmer (2005). We simulated a $p\times n$ data matrix **X** composed of n independent random samples from p dimensional multivariate Gaussian distribution $N_p(\mathbf{0}, \mathbf{\Sigma})$, where $\mathbf{\Sigma}$ is determined by a simulated concentration matrix $\mathbf{\Omega}$. We initialized $\mathbf{\Omega}$ by a $p \times p$ matrix with all elements being 0's. Given η, the proportion of null edges among all the $p(p-1)/2$ edges, we randomly selected $100(1-\eta)\%$ of the off-diagonal elements of $\mathbf{\Omega}$ and filled in values from uniform distribution on $[-1,1]$. To ensure that $\mathbf{\Omega}$ is a positive definite matrix, the diagonal elements of $\mathbf{\Omega}$ were filled by column-wise sums of absolute values plus a small constant. Finally $\mathbf{\Sigma}$ was calculated by $\texttt{scale}(\mathbf{\Omega}^{-1})$.

Let $|\mathbf{E}|$ be the number of edges in set **E**. The Gaussian simulation settings are

1. $p = 50$, $n = 100$, and $|\mathbf{E}| = 45, 55, 65, 75$ $(1-\eta \approx 0.037, 0.045, 0.053, 0.061)$
2. $p = 200$, $n = 100$, and $|\mathbf{E}| = 160, 200, 220, 240$ $(1-\eta \approx 0.008, 0.01, 0.011, 0.012)$.

   Because our proposed method does not require **X** to follow multivariate normal distribution, we assumed that each column of **X** is independently following multivariate central t-distribution with degrees of freedom 2 after specifying $\mathbf{\Sigma}$. The t simulation setting is

3. $p = 194$, $n = 75$, and $|\mathbf{E}| = 160$ ($1 - \eta \approx 0.008$)

This simulation setting has the median level of dimensionality and sparsity of the gene clusters studied in our real data example in section 3.3.

We first evaluated the accuracy of partial correlation graph using ROC curves, and compared our method with one of the most widely used method for partial correlation matrix estimation, the Graphic Lasso (GLasso) (Friedman et al., 2008). For our method, the ROC curve is drawn by selecting optimal ridge parameter λ using 10-fold cross-validation, and then calculating sensitivity and specificity across different p-value cutoffs α. This approach is less ideal than the two-grid search of λ and α, but the ROC curve more smooth and is easier to interpret since the sensitivity is a monotone function of α. For GLasso, we used the implementation in R package GLasso, and the sensitivity and specificity were calculated across different values of the tuning parameter κ ranging from 0.5 to 1000. For each p-value cutoff for our method and κ for GLasso, 1-specificity and sensitivity are averaged across 100 simulation data. As shown in Figure 2, our method has uniformly better sensitivity and specificity than the GLasso in estimating network structure.

These simulation studies also demonstrate that our density estimates of the test statistics ψ (the Fisher's Z-transformation of ridge estimates of partial correlations) fit the observed distribution well (Figure 3). To estimate the sparsity η (the proportion of partial correlations that are zero), there are two possible methods: (1) the proportion of zero partial correlations obtained from our thresholding step; and (2) an estimate of the upper bound of η by (Efron, 2004) (see Supplementary Materials Section C for details). For one typical simulation in the low dimension setting with true sparsity 0.963 (Figure 3 (a)), the estimated null density is $\hat{f}_0 = N(-0.002, 0.025^2)$ and the estimated sparsity level is $\hat{\eta} = 0.960$ (Efron's upper bound is 0.981). For one typical simulation in the high dimension case with true sparsity 0.992 (Figure 3 (b)), the estimated null density is $\hat{f}_0 = N(-0.0002, 0.011^2)$ and the estimated sparsity level is $\hat{\eta} = 0.988$ (Efron's upper bound is 0.995). An alternative choice to threshold partial correlation estimates is the hypothesis testing approach by Schäfer and Strimmer (2005), which is an EM algorithm assuming null and alternative distributions are Hotelling distribution and uniform distribution, respectively. However the density curves of null and mixture distributions of their method do not fit the empirical distribution well and thus we do not further pursue it here (Supplementary Materials Section E).

We evaluated the accuracy of partial correlation coefficient estimates using sum squared error (SSE), and compared our method with GLasso and the method of Schäfer et al. (2005), where partial correlation matrix was estimated by inverse of an optimal estimate of correlation matrix (implemented by function `pcor.shrink` of R package `corpcor`). Given the set of vertices $\mathbf{\Gamma} = \{1, 2, \ldots, p\}$, The SSE was calculated as

$$L(\mathbf{R}, \hat{\mathbf{R}}) = \sum_{a \neq b \in \mathbf{\Gamma}} (\hat{\rho}_{ab} - \rho_{ab})^2, \quad (17)$$

where $\hat{\mathbf{R}} = [\hat{\rho}_{ab}]_{p \times p}$ was the estimates of $\mathbf{R}$. The mean values of SSE from 100 replicates of each simulation setting were reported. As shown in Figure 4, our method consistently show better performance than GLasso and the method of Schäfer et al. (2005).

The ROC curves and SSE results of other simulation settings for our method and GLasso are displayed in Figures S7–S12 (Supplementary Materials Section F). As number of edges $|\mathbf{E}|$ increases, the performance of our method deteriorates while the performance of GLasso improves. For denser graph, the assumption that the null density match with the mixture density at the central part may be violated, and thus the estimate of null density may be biased.

3.3 Application

We applied our method to estimate the partial correlation graph of the expression of 6178 genes from yeast cell cycle data (Spellman et al., 1998). The gene expression data were downloaded from http://genome-www.stanford.edu/cellcycle/data/rawdata/. After removing the samples with more than 20% missing values, 75 samples remained for further analysis. We imputed the remaining missing values of the expression data using nearest neighbor averaging. Then the expression data of each sample were normalized by quantile normalization. In this analysis we did not account for the time-dependent nature of the data and treated the expression of each gene across 75 samples as independent observations.

Denote the observed gene expression data as a matrix $\mathbf{X}$ of dimension 6178×75. Each gene is a variable, and thus $\mathbf{\Gamma}$ is {1, …, 6178}. We first grouped the 6178 genes into h clusters. Let $\mathbf{C}_i$ be the genes belonging to the i-th cluster, then $\sum_{i=1}^{h}|\mathbf{C}_i|=6178$. We separately constructed the partial correlation graph within each cluster. The graph of ith cluster is denoted by $\mathbf{G}_{\mathbf{C}_i} = (\mathbf{C}_i, \mathbf{E}_{\mathbf{C}_i})$ for $i = 1, \ldots, h$. We assumed the genes from different clusters were independent, so that the edge set $\mathbf{E}$ of the whole graph was estimated by $\hat{\mathbf{E}}=\cup_{i=1}^{h}\hat{\mathbf{E}}_{\mathbf{C}_i}$. Specifically, we clustered the 6178 genes using hierarchical clustering with Ward's minimum variance method and the distance between two genes a and b was defined as $1-|\hat{\rho}_{ab|\emptyset}|$ where $\hat{\rho}_{ab|\emptyset}$ denoted marginal Pearson correlation. We chose the number of clusters to be 25 based on the *gap* statistic (Tibshirani et al., 2001). Cluster sizes varied from 32 to 1370, with 25 percentile, median, and 75 percentile being 139, 194, and 254, respectively. Among all the 19,080,753 gene pairs of the 6178 genes, 1,556,154 belonged to the same cluster, and hence could be connected based on the partial correlation graph estimates.

The density curves estimated by our method fit the observed data well. For example, Figure S13 (Supplementary Materials Section F) shows the results for one cluster including 124 genes. Here the null distribution is $N(0.003, 0.028^2)$, and the density of the mixture distribution is estimated using the 5 degrees of polynomial Poisson regression.

We compared the performance of our method with the GLasso (Friedman et al., 2008). We chose the tuning parameter κ of GLasso based on the extended Bayesian information criterion (BIC) (Foygel and Drton, 2010):

$$\mathrm{BIC}_{\gamma}(\kappa)=-\log|\hat{\mathbf{\Omega}}(\kappa)|+\mathrm{tr}(\hat{\mathbf{\Omega}}(\kappa)\mathbf{S})+\frac{1}{\mathrm{n}}|\hat{\mathbf{E}}(\kappa)|\log\mathrm{n}+\frac{4}{\mathrm{n}}|\hat{\mathbf{E}}(\kappa)|\gamma\log\mathrm{p}, \quad (18)$$

where $\hat{\mathbf{\Omega}}(\kappa)$ was the estimate of the inverse covariance matrix using GLasso with tuning parameter κ, $\hat{\mathbf{E}}(\kappa)$ was the edge set obtained from $\hat{\mathbf{\Omega}}(\kappa)$, and $\gamma \in [0, 1]$ is a tuning parameter

of the extended BIC. If $\gamma = 0$, the classical BIC used in Yuan and Lin (2007) was recovered. Given a fixed γ value, we applied GLasso with tuning parameter selected by the extended BIC (18) to construct the partial correlation graphs for all 25 clusters. An exception is that for two clusters with low dimension such that $p < n/2$, we always used the classical BIC by setting $\gamma = 0$. Different choices of γ led to the different model selection results.

The estimates of partial correlation graphs were evaluated by comparing the edge set $\hat{\mathbf{E}}$ with yeast protein-protein interaction database at http://thebiogrid.org/download.php. Table 1 displays the number of directed edges, the number of undirected edges (after omitting the directions) and the number of vertices in each of 20 protein-protein interaction dataset. We considered two genes are truly connected if they belonged to the same cluster and the corresponding proteins had interaction according to at least one of the protein-protein interaction datasets. Among 1,556,154 gene pairs belonging to the same cluster, 9,382 were connected and 1,546,772 were not connected. Many gene pairs may be connected according to their partial correlation, but may not be connected by this standard of protein protein interaction. Therefore by defining true connections by protein-protein interaction, we would over-estimate false discovery rate. However, such over-estimation does not interfere with comparison across different methods. Given this imperfect, but biologically meaningful definition of true/false connections, we evaluated our method and GLasso by ROC curves. The ROC curve of our method was generated across different p-value cutoffs α and the ROC curve of the GLasso was generated across different γ values in the extended BIC. Our method had uniformly higher sensitivity and specificity than the GLasso (Figure 5).

4. Discussion

Motivated by the gene co-expression network estimation problem, we have developed a new framework for estimation and statistical inference of partial correlation matrix. Both simulation and real data analysis have demonstrated the effectiveness of our method. For real data analysis where p is much larger than n, we cluster the genes and then estimate partial correlation matrix within each cluster. This is based on an reasonable assumption that the partial correlation matrix of gene expression has a block diagonal structure. We used the hierarchical clustering to group genes. There are many other clustering method available Monti et al. (2003); Zhang et al. (2005), though a careful study of which clustering method can better identify the block diagonal structure is beyond the scope of this paper. Our method does not require multivariate Gaussian distribution assumption. However, without this assumption, partial correlation being zero may not imply the two variables are conditionally independent with each other.

The first step of our method, ridge estimation of partial correlation matrix is closely related with Schäfer et al. (2005)'s method. However the objective of our method is to estimate partial correlation matrix while Schäfer et al. (2005)'s method is designed to estimate correlation matrix, and the optimal estimate of correlation matrix does not guarantee the optimal estimate of partial correlation matrix, as shown in our simulation studies. In terms of implementation, Schäfer et al. (2005) estimated covariance matrix in one step using a single tuning parameter, which is selected to minimize the squared loss of correlation matrix estimate. In contrast, we estimate partial correlation matrix in three steps with two tuning

parameters, which are selected by a two-grid search to minimize the squared loss of partial correlation matrix estimate.

While the work in the field of partial correlation matrix estimation is moving toward more sophisticated penalized estimation, our work shows that a conceptually simple approach of ridge estimation + thresholding + reestimation can deliver surprisingly good results. One key of the success of our method is that we borrow information across variables to reduce estimation variance, while scarifying estimation bias. This point is more clear when comparing our method with neighborhood selection by ridge regression. In neighborhood selection, different tuning parameter λ's will be used for different variables, which will reduce the bias of ridge estimation, but increases variance. In contrast, our method borrows information across all the variables by utilizing a common tuning parameter λ across all variables, and this approach increase the bias of the partial correlation estimate, but reduces the variance. The reduced variance leads to better discrimination of zero and non-zero parietal correlations in the thresholding step, and the inated bias can be corrected in the final reestimation step.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We are grateful for constructive comments from the anonymous reviewers, and the associate editor. These comments lead to significant improvement of the work presented in this paper. We would like to acknowledge the support from EPA grant RD-83382501, and NIH grant P01 CA142538-03, R01 CA149569-05, and R01 MH101819-01.

References

Anderson, T. An introduction to multivariate statistical analysis. Wiley Series in Probability and Statistics; 2003.

Banerjee O, El Ghaoui L, d'Aspremont A. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. The Journal of Machine Learning Research. 2008; 9:485–516.

Castelo R, Roverato A. A robust procedure for gaussian graphical model search from microarray data with p larger than n. The Journal of Machine Learning Research. 2006; 7:2621–2650.

Christensen, R. Plane Answers to Complex Questions: The Theory of Linear Models. Springer; 2002.

de Jong S, Boks M, Fuller T, Strengman E, Janson E, de Kovel C, Ori A, Vi N, Mulder F, Blom J, et al. A gene co-expression network in whole blood of schizophrenia patients is independent of antipsychotic-use and enriched for brain-expressed genes. PloS one. 2012; 7:e39498. [PubMed: 22761806]

Dempster A. Covariance selection. Biometrics pages. 1972:157–175.

Efron B. Large-scale simultaneous hypothesis testing. Journal of the American Statistical Association. 2004; 99:96–104.

Fan J, Feng Y, Wu Y. Network exploration via the adaptive lasso and scad penalties. The Annals of Applied Statistics. 2009; 3:521. [PubMed: 21643444]

Foygel R, Drton M. Extended bayesian information criteria for gaussian graphical models. arXiv preprint arXiv:1011.6640. 2010

Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical lasso. Biostatistics. 2008; 9:432–441. [PubMed: 18079126]

Hoerl A, Kennard R. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics. 1970; 12:55–67.

Højsgaard S, Lauritzen S. Graphical gaussian models with edge and vertex symmetries. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2008; 70:1005–1027.

Lauritzen, S. Graphical models. Vol. ume 17. USA: Oxford University Press; 1996.

Lv J, Fan Y. A unified approach to model selection and sparse recovery using regularized least squares. The Annals of Statistics. 2009; 37:3498–3528.

Magwene P, Kim J, et al. Estimating genomic coexpression networks using first-order conditional independence. Genome Biol. 2004; 5:R100. [PubMed: 15575966]

Meinshausen N, Bühlmann P. High-dimensional graphs and variable selection with the lasso. The Annals of Statistics. 2006; 34:1436–1462.

Monti S, Tamayo P, Mesirov J, Golub T. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. Machine learning. 2003; 52:91–118.

Rothman A, Bickel P, Levina E, Zhu J. Sparse permutation invariant covariance estimation. Electronic Journal of Statistics. 2008; 2:494–515.

Schäfer J, Strimmer K. An empirical bayes approach to inferring large-scale gene association networks. Bioinformatics. 2005; 21:754–764. [PubMed: 15479708]

Schäfer J, Strimmer K, et al. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Statistical applications in genetics and molecular biology. 2005; 4:32.

Schott, J. Matrix analysis for statistics. Wiley; 2005.

Speed T, Kiiveri H. Gaussian markov distributions over finite graphs. The Annals of Statistics. 1986; 14:138–150.

Spellman P, Sherlock G, Zhang M, Iyer V, Anders K, Eisen M, Brown P, Botstein D, Futcher B. Comprehensive identification of cell cycle–regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. Molecular biology of the cell. 1998; 9:3273–3297. [PubMed: 9843569]

Stuart J, Segal E, Koller D, Kim S. A gene-coexpression network for global discovery of conserved genetic modules. Science. 2003; 302:249–255. [PubMed: 12934013]

Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a data set via the gap statistic. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2001; 63:411–423.

Wille A, Bühlmann P. Low-order conditional independence graphs for inferring genetic networks. Statistical applications in genetics and molecular biology. 2006; 5:1–32.

Wille A, Zimmermann P, Vranová E, Fürholz A, Laule O, Bleuler S, Hennig L, Prelic A, Von Rohr P, Thiele L, et al. Sparse graphical gaussian modeling of the isoprenoid gene network in arabidopsis thaliana. Genome Biol. 2004; 5:R92. [PubMed: 15535868]

Yuan M. High dimensional inverse covariance matrix estimation via linear programming. The Journal of Machine Learning Research. 2010; 99:2261–2286.

Yuan M, Lin Y. Model selection and estimation in the gaussian graphical model. Biometrika. 2007; 94:19–35.

Zhang B, Horvath S, et al. A general framework for weighted gene co-expression network analysis. Statistical applications in genetics and molecular biology. 2005; 4:1128.

Zhou S, Rütimann P, Xu M, Bühlmann P. High-dimensional covariance estimation based on gaussian graphical models. The Journal of Machine Learning Research. 2011; 12:2975–3026.

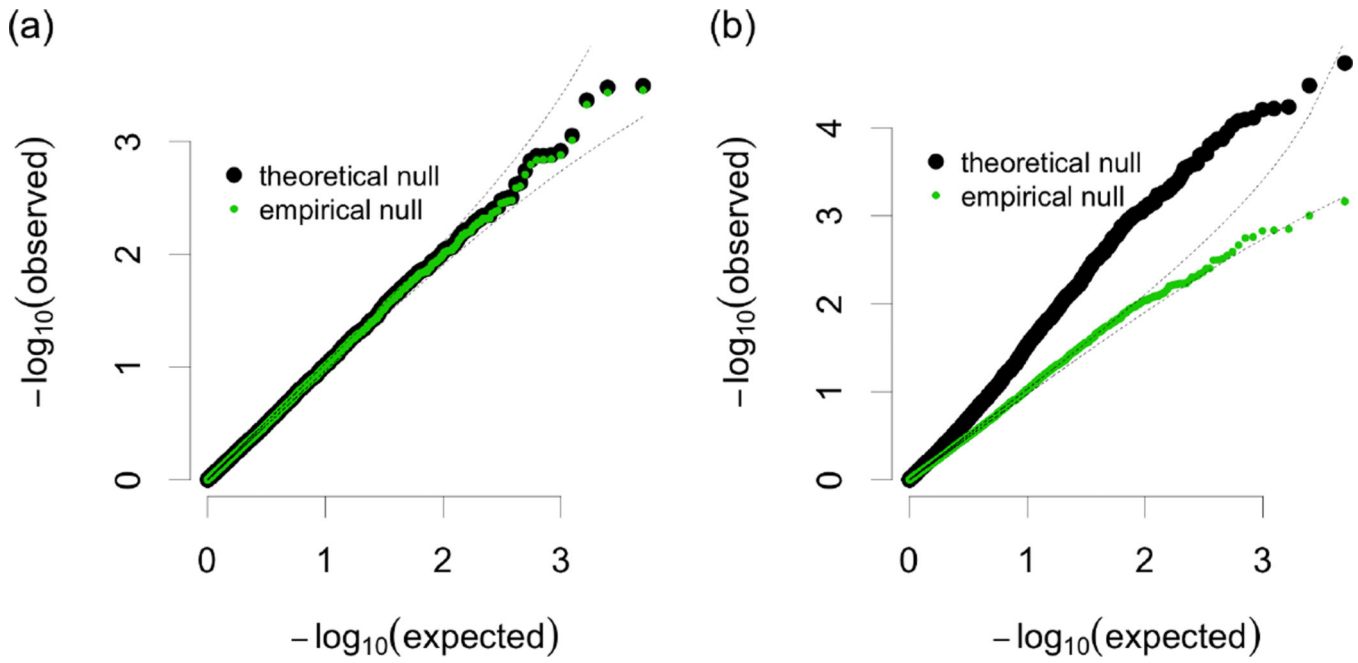

Figure 1.

QQ-plots for p-values calculated using theoretical null distribution (larger points) or empirical null distribution estimated by central matching method (smaller points) against the expected uniform distribution on [0,1]. (a) $p = 100$ and $n = 1000$, (b) $p = 100$ and $n = 110$. The dotted lines are the 90% confidence limits of the expected values.

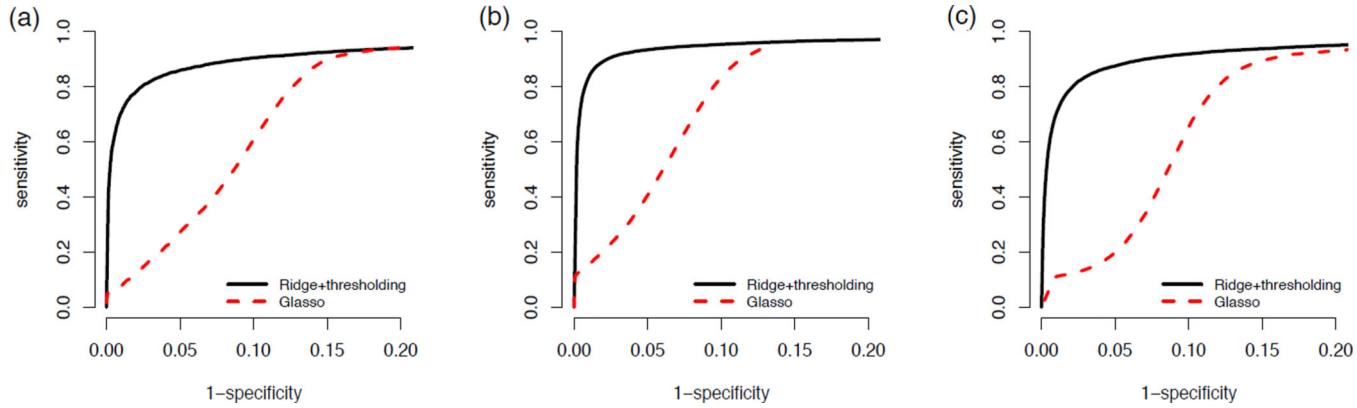

Figure 2.

The ROC curves for identifying zero entries of partial correlation matrix using our method or GLasso for three simulation settings: (a) Mutivariate Gaussian for $n = 100$, $p = 50$, and $|\mathbf{E}| = 45$. (b) Mutivariate Gaussian for $n = 100$, $p = 200$, and $|\mathbf{E}| = 160$. (c) Mutivariate t-distribution for $n = 75$, $p = 194$, and $|\mathbf{E}| = 160$.

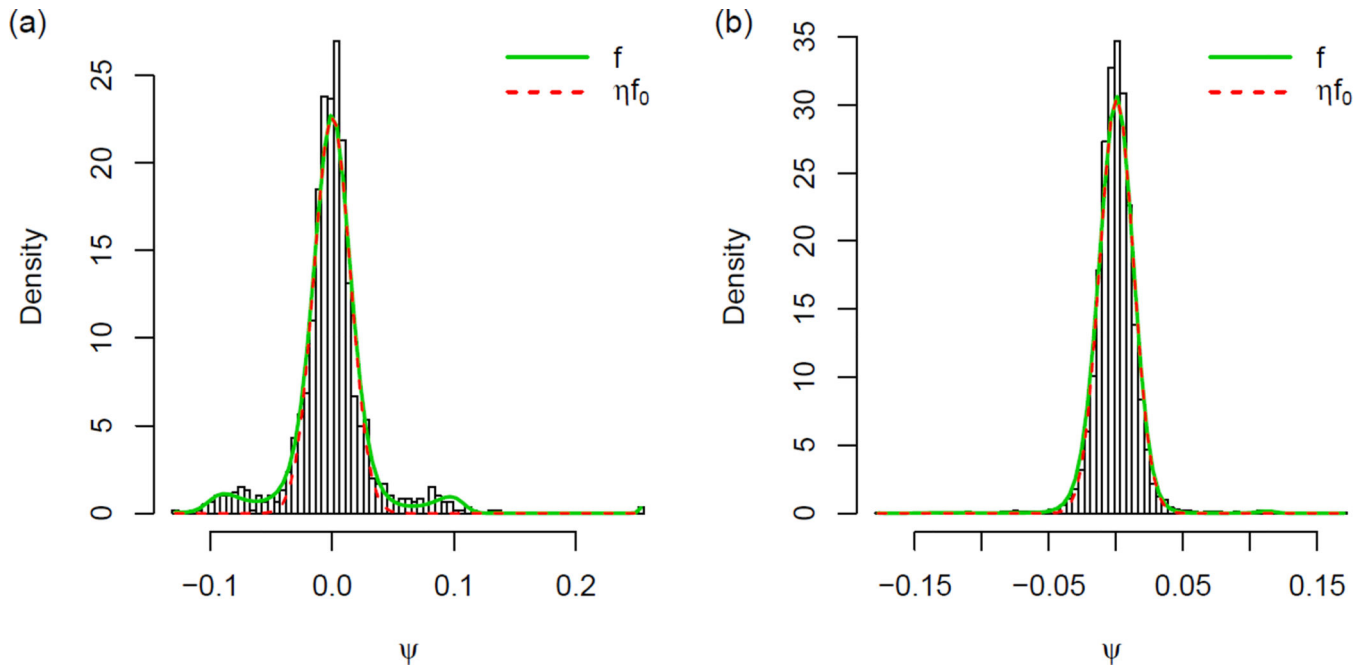

Figure 3.
Histograms of ψ values and density curves of the estimated null and the marginal density. (a) $p = 50$, $n = 100$ and $|\mathbf{E}| = 45$ (b) $p = 200$, $n = 100$ and $|\mathbf{E}| = 160$.

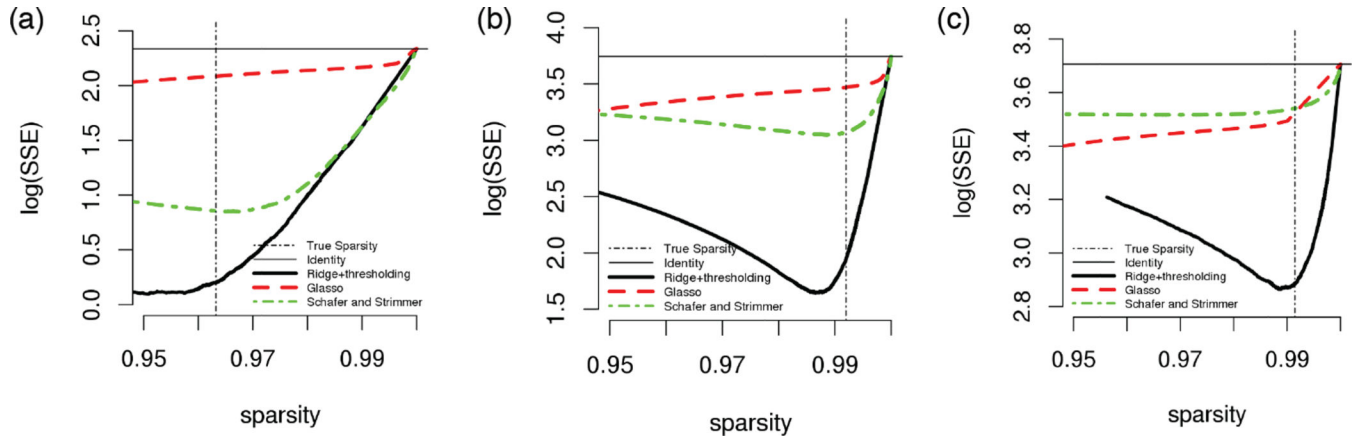

Figure 4.

Sparsity (proportion of zero partial correlation estimates) versus log(SSE) for three settings: (a) Gaussian, $n = 100$, $p = 50$, and $|\mathbf{E}| = 45$; (b) Gaussian $n = 100$, $p = 200$, and $|\mathbf{E}| = 160$; (c) t-distribution $n = 75$, $p = 194$, and $|\mathbf{E}| = 160$. The horizontal black line is log(SSE) values when a $p \times p$ identity matrix is used and the vertical black line indicates the sparsity of the true network.

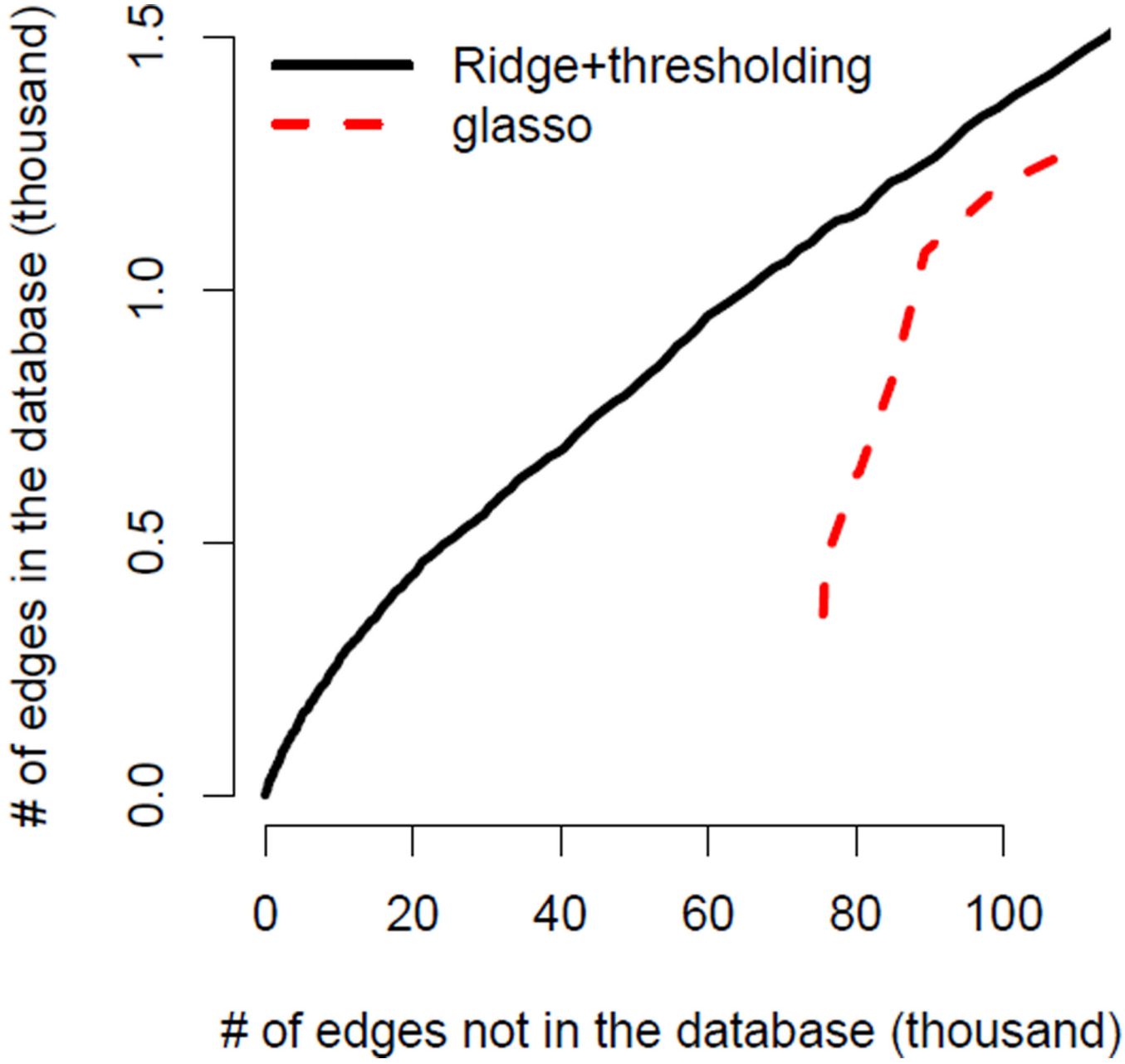

Figure 5.
Comparing our method (Ridge+thresholding) with GLasso in terms partial correlation graph estimation by ROC curves, while the underlying true connections are defined as gene pairs belonging the the same cluster and their proteins having protein-protein interaction.

Table 1

Summary of the protein-protein interaction database

ID	Experiment system (type)	no. of directed edges	no. of undirected edges	no. of vertices
1	Affinity Capture-MS (physical)	72767	42538	4613
2	Affinity Capture-Western (physical)	13105	7795	2727
3	Dosage Rescue (genetic)	4812	4022	2161
4	Reconstituted Complex (physical)	5110	3946	1988
5	Synthetic Lethality (genetic)	13870	10965	2915
6	Two-hybrid (physical)	13986	10827	3392
7	Biochemical Activity (physical)	5703	5220	1946
8	Co-crystal Structure (physical)	387	337	421
9	FRET (physical)	142	119	117
10	Protein-peptide (physical)	673	643	353
11	Co-localization (physical)	527	484	441
12	Affinity Capture-RNA (physical)	5895	5888	3702
13	Protein-RNA (physical)	408	399	377
14	PCA (physical)	5117	4845	1663
15	Co-purification (physical)	1675	1309	933
16	Co-fractionation (physical)	777	725	663
17	Dosage Lethality (genetic)	971	945	786
18	Phenotypic Enhancement (genetic)	6449	4803	2153
19	Phenotypic Suppression (genetic)	5287	3965	1729
20	Synthetic Haploinsufficiency (genetic)	262	262	262